CSCI 620 – Intro to Big Data

Assignment #7 – Frequent Itemset Mining

.Name: *Yuvraj Singh*

1) **Creating Popular_Movie_Actors**

Within the *assignment7.py* script, there is a method *def popular_movie_actors* that will create a table called Popular_Movie_Actors which selects information from the title_actor table where the title type is 'movie' and the avg rating is about a 5.

2) **Frequent Itemset Mining of Size One**

In the python script, there is a method that calculates the item sets that have a minimum support of five. The method *def item_set_size_one* will generate a table called 'L1' by counting the number of instances of each actor in the popular_movie_actors table.

3) **Frequent Itemset Mining of Size Two**

This section of the assignment is very similar to the previous section but instead, we join on popular_movie_actors even though we selected from it initially so we can obtain a second actor. To exclude pairs where we would have duplicates like (Bread, Butter) and (Butter, Bread), the query made sure that the next actor id was always greater than the first actor id selected. The method for this section is *def item_set_size_two*.

4) **Frequent Itemset Mining of Size Three**

Like the previous section, another join was added onto the query to select the third actor. The method for this section is *def item_set_size_three*.

5) Frequent Itemset Mining of Size N

In order to obtain any lattice, a method called *def item_set_size_n_query* will generate the queries to select the correct information to create lattice n. The method *def freq_item_set* utilizes the method *def item_set_size_n_*query to create each lattice up until the last empty lattice. What the method does is it first performs a select query to see how many rows will be in the lattice query, if the result is not 0, then it will create the lattice of size n. It will begin at one and continue until there cannot be any more lattices to create. When it creates a lattice, the script will print out the number of items in that lattice, and it will continue to increment n and create a new lattice until the empty lattice is reached. Once the empty lattice is found, it will create a new table called last_non_empty_lattice which is the last lattice that wasn't empty, and it will include the names of each actor.

```
There are 42806 items in L1
There are 13314 items in L2
There are 2297 items in L3
There are 867 items in L4
There are 426 items in L5
There are 125 items in L6
There are 13 items in L7
L8 is an empty latice
```